

Notes
Data Warehouse (DW) Technical Review and Planning Meeting
April 3, 2002
Albuquerque, NM

Meeting Goals, as stated by Russell Pittman, Director, ITSC:

- Summarize activities to date, ensure that all are on the same page in order to make decisions for next steps
- Determine how the current Pilot Data Warehouse (PDW) project would transition into next phase
 - How and when will PDW impact on existing processes and systems
- How to coordinate DW project with other key and related technology projects, i.e., Enterprise Master Person Index (EMPI), data transport, etc
 - Acknowledging that the current structures of data transformation and transport is left over from mainframe environment, what new tools can we use now
- Identify the impact of moving PDW into production environment: funds, staff, etc.

The overall goal of the meeting is to for all participants to have the opportunity to express their opinions on all aspects of planning and implementing a Data Warehouse.

The assumptions for all decision making include the following:

- Not to lose the functionality we currently have
- Need to live technologically within the mainframe and DB2 environment
- The data warehouse will physically reside in Albuquerque

8:30-9:30 AM

Stan Griffith provided an Overview of the Data Quality Action Team activities over the past year, including the planning and implementation of a Pilot Data Warehouse (PDW) (see presentation handouts).

Key discussion points from group q&a after the presentation:

- Would definitely plan to expand into other data areas besides current, particularly business
- The Data Warehouse (DW) would expect to contain “almost” raw data, i.e., with minimal transformation/cleansing between receipt and tabling, such as correct data formats.
- The DW would be expected to have multiple data marts where data transformation would take place depending on the program needs of the data mart owner. Data marts would include statistical (e.g., NPIRS), and clinical (Epi or Diabetes, GPRA or IHPES), etc.
- The intention is for Urban data to be included as well.
- Data will need to be continuously analyzed to identify specific types of data issues and possible solutions. For example, fields in RPMS may need to be changed.
- ITSC will need to communicate clearly with the field about the different components of the DW and the data marts. We must publicize to the field that the data will not be transformed for the DW and that Areas can have direct access to their data via a data mart. Must also

communicate with Areas regarding data quality and provide them with information about what needs to be corrected,

10:00 AM – 12:30 PM

Two presentations described in additional detail the technical and functional approaches to the Pilot Data Warehouse (PDW) and reporting. The third session was a presentation of options for next steps. (See slide handouts for further details).

Review of Pilot Data Warehouse: Stephanie Klepacki

- Data from 9 sites (three from three Areas) and one non-RPMS tribal site reporting; CHS data from NPIRS raw files
- PCC and patient registration data to PDW, run through BPX (formerly AIB)
- Overview for ETL (export, transform and load): source data to Staging db, near raw data; from Staging, moves through additional minor data cleansing to DW.
- Model includes parent tables with data that doesn't repeat, child tables for repeating data, and state tables for data that likely changes. Rejected data is sent to an error table.
- Model developed using exported data; industry standard models have data we don't need and don't have data we do need.

SAS Reporting and Analyses: Karen Carver

- Access will be handled by Java Script over the Web with security at all levels.
- Users of data marts will determine denominators and numerators for reporting, just as they do now.

DW Implementation Issues and Strategies: Mike Gomez

Seven implementation strategies were described that covered a wide range of options:

- Complete PDW development "as is," as described in Stephanie's presentation, as a proof of concept, and continue to maintain two production databases for NPIRS and ORYX.
- Expand existing PDW to another pilot (PDW-2), with no standards-based (HL7) data transport, in order to test PCC Export patch 6 and fine tune the export, transformation and load (ETL) process.
- Complete PDW as planned and implement a production data warehouse (DW-1), maintaining the IHS custom interface PCC Export patch 6 rather than implementing standards-based data transport (HL7).
- Revise PDW Strategy to implement standards-based data transport.
- Revise PDW Strategy to include other changes, but not to include standards-based data transport.
- Complete PDW as planned and implement a production data warehouse (DW-1) using standards-based data transport.
- Monitor VHA efforts to design and develop a Clinical Data Repository (CDR).

Substantial discussion took place about

- how and why to implement HL7;
- the potential role of the Generic Interface Engine (GIS) and/or the Cloverleaf Interface Engine;
- where some or all data transformation would take place in the various scenarios;
- positive and negative impacts of maintaining IHS custom interface rather than moving toward standards-based transport.

Afternoon Session

Rus Pittman facilitated the afternoon session by posing to the group a series of questions about various aspects of the data warehouse plan.

Is the PDW scalable from 10 to 400+ sites

- Are load rates scalable? Some slow load rates, need improvement. Load from source to Staging is quick; slower rates come from Staging to DW, because of checking the data and transformation. Currently 28 hours to load nine sites (approx 3 hours each) from Staging to DW. Options to be explored include using DB2 insert statements instead of the cursor system, and repartitioning the database into 4 partitions.
- Is the architecture scalable? Yes.
- Is hardware/disk space/etc. scalable? Differentiate between needs for initial loading and for routine production
- Impact of Patch 6: PCC Export patch 6 has not yet been implemented, will have to revise ETL and test. Patch 6 does not quite match PDW logic
- Eligibility data has not been loaded but should be included in DW – would we reload all data to include eligibility?
- Is the data model optimized for production? Flatter? Or too flat?

Data Mart Development

- What will it take to stand up a data mart? Lisa estimates 1 FTE for 1 to 3 months from requirements to completion. Big challenge is in building initial data sets.
- Will need implementation plan to include user training, tools, assistance with requirements, etc. It is anticipated that most data mart owners will use existing staff to assist them; some large data marts may have their own SAS or other analyst.
- Data marts can be in different applications (e.g., DB2, SAS, etc). Should provide 2-3 standards to users to discourage proliferation ORYX will continue to use SAS datasets.

Using HL7 from Sites to Interface Engine (Cloverleaf or GIS)

- Advantage to keeping current export procedure: some Areas manipulate local data first and/or prefer to send all SU data together. Options: 1) Interface Engine (IE) could send to Area and to DW simultaneously – HL7 can send updated data; 2) Areas would subscribe to Area data mart (DM).
- Do the HL7 triggers make a larger # of records go to NPIRS? What is impact? HL7 triggers is a separate issue from exporting data in HL7 format. Handling of re-exports being drafted.

- If the current process is maintained, the export process has to be modified (patch) when data needs change. With HL7, it's a table change in GIS or IE.
- Currently GIS has mapped most of outpatient, some of inpatient
- Downside to HL7 and Cloverleaf: lag between ANSI certification and into upgraded software.
- With HL7, there is a security segment, PKI in the message for authentication.
- Data is currently being moved into an FTP directly where it can be looked at by others before being moved. With HL7 could continue as FTP or open TCP/IP and receiving port to be sent so not sitting – automatic notifications and paging when links are down
- What processing do we lose that currently exists in BXP? Aggregates data; some reporting to Areas. Cloverleaf can provide certain types of information to Areas also.
- What about non RPMS sites? Most are using one of three systems: Health Pro, Medical Manager, or HealthTek (?). Trying to get HealthPro sites to use PCC and export data. May be better use of our resources to put time into assisting Tribes to move to HL7 and work with the key vendors.
- Are there any legislative requirements for IHS to use HL7? Within 6 months on the billing side only. However, HL7 interface initiatives are proliferating in the field.

Using HL7 from Cloverleaf Interface Engine (IE) to Data Warehouse

- Test plan for HL7 is currently from IE to NPIRS. Should that change?
- Area gets separate files from SU and currently sends as one file. NPIRS requires certain file names. With HL7, this info is part of the header information.
- What will it take for NPIRS to accept HL7? An interface and parser are needed.
- Can or should IE perform the ETL and load the DW? Or is the IE just doing the routing: convert non HL7 to BXP or BXP to HL7. What, if any, changes would be needed to ETL to move into IE. If 2 man months are already invested in the ETL within the DW, would it be a minor or major investment to move ET to the IE.
- Would HL7 go back and process historical data (through 1997) or just begin HL7 at a point in the future? We could continue to load the DW while changing to HL7

PDW Completion

- What is reasonable timeframe for completing PDW: evaluate, identify and fix? 4-6 months to do following
 - SAS is 20-25% completed
 - Eligibility data and other mods to model
 - Testing on load time
 - Transition from IBM consulting staff to internal resources
 - CPT
 - Non RPMS site data
 - Already have written ETL for PCC, Denrun, CHSFI

Scenario Proposed by Rus Pittman

- Existing PDW finishes its run, with some enhancements, e.g., Eligibility, model
- DW ignores non-RPMS – let IE handle that interface
- Ignore any DW ETL that has not been written already
- Existing export mechanism to IE
- IE will provide v. 6 “like” file to DW
- Move Data Transport test from NPIRS to DW

Near Term Actions:

- Need calculation on how long it will take to load all data
- Identify options to improve load speed, e.g., partitioning the database
- Look at logic for data checking from Staging to DW that is production heavy and determine what could be eliminated. What is the frequency of the problems that are being checked for?
- Identify impact of data loading on other systems using same box, and vice versa – is additional space or another box needed for the initial data load
- Evaluate existing PDW structure for Data Warehouse use, id model enhancements
- Differentiate between initial PDW load and production
- Identify generic steps and schedule for identifying and implementing data marts
- Process for new data elements for Metadata Registry (Jim McCain)